Custodial Institutions Agency
Ministry of Justice and Security

Interrater and Intrarater Reliability of the Violent Extremism Risk Assessment tool

Annemaryn de Bruin
Nils Duits
Maaike Kempes
Merel Prinsen

Research report 2022-1

Science and Education, NIFP

Annemaryn A. de Bruin, MSc. Junior researcher at the Netherlands Institute of Forensic Psychiatry and Psychology (NIFP) of the Dutch Ministry of Justice and Security.

Nils Duits, MD, PhD. Forensic child- and adolescent psychiatrist & senior researcher at the Netherlands Institute of Forensic Psychiatry and Psychology (NIFP) of the Dutch Ministry of Justice and Security. Co-author and supervisor-trainer of the Violent Extremism Risk Assessment tool (VERA-2R). Former research coordinator of the VERAPRO2EU Project.

Prof. dr. Maaike Kempes. Head of Department of Science and Education, Netherlands Institute of Psychiatry and Psychology (NIFP) of the Dutch Ministry of Justice and Security.

Merel Prinsen, MA, LLM, PhD. Research coordinator of the VERAPRO2EU Project.

Acknowledgements:
This work was supported by the European Commission under Grant (Call JUST-JCOO-AG-2020, Project 101007383). We want to thank them for their financial support. Furthermore, we would like to thank the Dutch Public Prosecution Service for providing the judicial files.

Disclosure Statement:
No potential financial interest or benefit was reported by the authors.

Correspondence:
Correspondence should be addressed to: Annemaryn A. de Bruin, Department of Science and Education, Netherlands Institute of Forensic Psychiatry and Psychology (NIFP), Dutch Ministry of Justice and Security, Utrecht 3511 EW, The Netherlands (email: annemaryndebruin@gmail.com).

Data Availability Statement:
The data are not publicly available due to its sensitive nature.

Co-funded by the European Union's Justice Programme (2014–2020)

Table of content

1. Abstract

The Violent Extremism Risk Assessment - Version 2 Revised (VERA-2R) is an evidence-based structured professional judgement (SPJ) tool for ideologically motivated violence. Use of the tool can help professionals in risk assessment and risk management of terrorists and violent extremists. It is important that the tool leads to reliable and valid risk assessments. Therefore, we aimed to establish the reliability of the VERA-2R, focusing on interrater- and intrarater reliability. In order to do so, trained researchers assessed a Dutch sample of convicted terrorist offenders, respectively of 30 cases and 33 cases, on the basis of extensive judicial files. In general, the average amount of agreement on the indicators and structured professional risk judgements can be classified as good to excellent, for both the interrater- and intrarater reliability. However, six indicators were found to have low reliability. In addition to clarifying the reliability of the VERA-2R, this study also showed how the interrater- and intrarater reliability of a SPJ tool can be investigated with trained assessors based on judicial files. This can be of added value, because existing reliability studies often use case vignettes, have small sample sizes and/or do not include a stringent training program. However, in order to develop a more reliable and valid VERA-2R tool, the remaining psychometric properties of the tool must be investigated.

2. Introduction

Although many definitions and types of terrorism exist, one can define terrorism in a more general way as the threat, preparation, or perpetration of serious violence based on ideological motives against people, or deeds aimed at causing socially disruptive material damage with the goal to cause social change, to instil fear among the population or to influence political decision-making (National Coordinator for Security and Counterterrorism, 2016). While no clear distinction between violent extremism and terrorism has fully evolved, violent extremism can be described as the beliefs and actions of people who support or use violence to achieve ideological, religious or political goals (Schmid, 2013; Striegher, 2015). Violent extremism and terrorism can have a wide range of detrimental consequences for society, including, among other things, the loss of human life, material damage, emotional impact, not to mention the damage to the democratic process and prevailing legal order. Individuals who are imprisoned for a violent extremist or terrorist offence, as well as prisoners who are radicalized in prison, pose a serious security threat, both during their imprisonment and after their release (Europol, 2022). Therefore, a more evidence-based professional approach to violent extremism and terrorism risk assessment and risk management is urgently needed (United Nations, 2018). Determining the psychometric properties of the currently available risk assessment tools for ideologically motivated violence is thus of utmost importance. In the present study, we seek to obtain greater insight into the reliability of the Violent Extremism Risk Assessment - Version 2 Revised (VERA-2R), focusing on the interrater- and intrarater reliability.

Violent Extremism and Terrorism Risk Assessment and VERA-2R

Evidence-based violence risk assessment can be defined as the process of collecting information about individuals, in a way that adheres to, and is guided by, the available scientific and professional knowledge-base, both for the purposes of understanding whether individuals constitute a risk of engaging in violent behavior in the near future and determining which subsequent actions should be taken to prevent this violence from occurring (Hart, 2009). Evidence-based risk assessments can inform risk management strategies and interventions by identifying the possible risk scenarios (Hart & Logan, 2011). Furthermore, they can help to ensure transparent decision-making, avoid recurring decision-making errors, and enhance the level of understanding both within and across multidisciplinary teams (Helmus & Thornton, 2015; Yang, Wong, & Coid, 2010). Traditionally, researchers and practitioners have distinguished between three generations of risk assessment methods: (1) unstructured clinical judgements, which are risk assessments that are based solely on clinicians' experience and knowledge (Roychowdhury & Adshead, 2014); (2) the actuarial method, which involves using a fixed algorithm to combine evidence-based indicators into a final risk judgement (Hart & Logan, 2011); (3) structured professional judgement (SPJ), which combines empirical knowledge with professional judgement (Hart & Logan, 2011). In order to arrive at a final structured professional risk judgement, assessors must use their professional judgement to integrate, combine and weigh all the relevant information and data related to the evidence-based indicators (Guy, Packer, & Warnken, 2012). Scientific experts consider SPJ to be the most suitable method for assessing the risk of ideologically motivated individuals (Pressman, 2009; Monahan, 2012; Sarma, 2017). Given that prior analyses have demonstrated that most of the

currently available risk assessment instruments for general violence are not relevant to the idiosyncrasies of terrorists and violent extremists, the need for a specialized SPJ instrument for ideologically motivated violence emerged (Pressman, 2009). This subsequently led to the development of different risk assessment tools for violent extremism and terrorism (Sarma, 2017).

The VERA was the first specialized tool for conducting individual risk assessments for terrorists and violent extremists (Pressman, 2009; Pressman, Duits, Rinne, & Flockton, 2018). In response to feedback from terrorism experts, the VERA was subsequently revised and renamed the Violent Extremism Risk Assessment - Version 2 (VERA-2) (Pressman et al., 2018; Pressman & Flockton, 2012). In 2018, the most recent version of the instrument, identified as the Violent Extremism Risk Assessment - Version 2 Revised (VERA-2R), became available. This version incorporated several further revisions and improvements based on additional research into the indicators associated with violent extremism and terrorism (Pressman & Duits, 2019; Pressman et al., 2018).

The VERA-2R can be used to establish the risk status for individuals who have been accused, arrested or convicted of a violent extremist or terrorist offence (Pressman et al., 2018). Adhering to the SPJ methodology, the VERA-2R acknowledges that the weighting of the indicators should not be defined beforehand, due to the fact that the relevance of the indicators may vary depending on the specific context of the individual (Pressman et al., 2018). Therefore, professional judgement must be exercised to integrate, combine and weigh all the relevant information and data related to the indicators (Guy et al., 2012). Based on the resulting final structured professional risk judgement, different risk scenarios must be formulated along with a corresponding risk management strategy for each of these scenarios (Douglas et al., 2014; Logan, 2017).

The VERA-2R is widely used by trained professionals, both within and outside Europe, to assist with decision-making within various stages of the criminal justice process (Van der Heide, Van der Zwan, & Van Leyenhorst, 2019). In pre-trial settings, the VERA-2R is used by probation officers, forensic psychiatrists and forensic psychologists to improve the risk recommendations they proffer to the court (Duits, Rinne, & Van Leyenhorst, 2017). In post-trial settings, the VERA-2R facilitates a tailor-made approach and differentiated placement policy (Duits et al., 2017), supports decisions about the continuation of intervention and/or rehabilitation programs, helps to determine whether prisoners are able to be released on parole, and is used to establish the risk that the persons under supervision will commit a violent extremist or terrorist offence in the future (Pressman et al., 2018).

Psychometric Properties of the VERA-2R

Risk assessments play a significant role in terms of combatting violent extremism and terrorism (European Commission, 2020; United Nations, 2018). Therefore, it is of importance that the risk assessment tools for ideologically motivated violence provide reliable and valid risk assessments. Reliability pertains to the extent to which a measurement is stable, consistent, predictable, accurate and free from random error (Groth-Marnet, 2009). Validity concerns the extent to which an instrument measures what it purports to measure (Field, 2005).

Despite the importance of validation, due to the relative dearth of thorough scientific research, there remains scarce knowledge about the reliability and validity of risk assessment instruments (Hartling et al., 2012). With respect to the VERA-2R, professionals have reached consensus on the face validity and content validity (Pressman et al., 2018). Face validity refers to the degree to which an instrument creates the impression that it encompasses the entirety of the concept that it claims to measure (Holden, 2010), while content validity can be defined as the degree to which an instrument adequately represents all the relevant facets of a given construct (Gyldmark & Morrison, 2001).

Interrater reliability pertains to the degree to which two or more observers independently score the same ratings for the feature that is being observed or measured (Multon & Coleman, 2018). High interrater reliability is important, since structured professional risk judgements serve as the basis for important decisions in the criminal justice process (Van der Heide et al., 2019), and therefore should be independent of the observers or professional assessors (Jonsson & Svingby, 2007).

High interrater reliability has already been demonstrated for the VERA (Beardsley & Beech, 2013). However, since the VERA-2R incorporates revisions that may affect the interrater reliability, this result cannot be generalized to the VERA-2R. In addition to this, Beardsley and Beech (2013) used case vignettes, had a small sample size and failed to include a stringent training program, which may have impacted upon their findings. In light of the above, further research is needed to establish the interrater reliability of the VERA-2R.

Previous research has not established the intrarater reliability of the instrument yet. Intrarater reliability refers to "the extent to which an assessor, reusing the same instrument, consistently assigns the same ratings over time while examining a single set of data" (Belur, Tompson, Thornton, & Simon, 2021). A high level of intrarater reliability is a prerequisite for risk assessment instruments, insofar as it indicates that the tool can measure a constant phenomenon in the same way over time (Hopkins, 2000), and therefore can be used to identify changes in risk. This is important, since the dynamic nature of the process of radicalization to ideologically motivated violence and vice versa requires repeated VERA-2R risk assessments (Pressman et al., 2018).

The term intrarater reliability is sometimes wrongly used interchangeably with the term test-retest reliability (Holmefur, Aarts, Hoare, & Krumlinde-Sundholm, 2009). However, there is a significant difference between the two. While intrarater reliability refers to the agreement between repeated observations of the same test session, test-retest reliability includes two different test sessions. Test-retest reliability therefore inevitably includes intrarater error (Holmefur et al., 2009). Since our research design includes static judicial files which will be examined twice by the same assessor, we will focus on the intrarater reliability.

Interrater and Intrarater Reliability of the VERA-2R

This study constitutes the first part of an extensive validation project and provides insight into the interrater and intrarater reliability of the VERA-2R. We hypothesize that the VERA-2R will have high interrater reliability, since the study of Beardsley and Beech (2013) demonstrated high interrater reliability for the VERA. In addition to this, given that the study of Beardsley and Beech (2013) provided a first indication that the VERA incorporates clearly expressed indicators and encoding rules, we hypothesize to find high intrarater reliability.

3. Methods Interrater Reliability

Assessors and Cases

The assessors were two Dutch researchers (one male, one female) with a bachelor and master degree in psychology and/or criminology. The researchers were employed by the Netherlands Institute of Forensic Psychiatry and Psychology (NIFP) of the Dutch Ministry of Justice and Security. The assessors took part in a two-day training course to obtain an in-depth understanding of the instrument and to acquire experience in applying the VERA-2R indicators and forming structured professional risk judgements.

To assess the interrater reliability of the VERA-2R, both assessors independently rated a sample of 30 terrorist cases based on extensive judicial files. These files were provided by the Dutch Public Prosecution Service and included a forensic mental health (FMH) assessment, a probation report, a transcript of the verdict, a police report, a criminal record and/or information from intelligence services.

The sample comprised 24 men and 6 women who were convicted of terrorist offences in the Netherlands between 2012 and 2019. The subjects' ages ranged from 15 – 47 at the time of their terrorist act (M_{age} = 25.47, *SD* = 7.99). Ninety percent of the subjects were from a migrant background, 3.3% of the subjects had no migration background, and for 6.7% the background was unknown. Most of the subjects from a migrant background had parents who were born in Morocco or were born there themselves (48.1%), followed by persons with a Turkish background (18.5%). Consequently, the sample appears to be representative of the target population with respect to age, gender and migration background, since similar descriptive statistics were found in previous studies on Dutch jihadists (Bakker & De Bont, 2016; Weenink, 2019).

Materials

The Violent Extremism Risk Assessment – Version 2 Revised (VERA-2R) contains 34 risk and protective indicators specifically related to the risk of violent extremism and terrorism (Pressman et al., 2018). The VERA-2R indicators are divided into five domains: Beliefs, Attitudes and Ideology (BA), Social Context and Intention (SCI), History, Action and Capacity (HAC), Commitment and Motivation (CM), and Protective and risk-mitigating indicators (P). The scientific basis for each indicator is explained, along with the underlying criteria for the three rating levels: low, moderate or high. A risk indicator is rated as 'low' if the risk-promoting indicator characteristics are objectively not present, as 'moderate' if the risk-promoting indicator characteristics are present to a specified level, and as 'high' if the risk-promoting indicator characteristics are clearly present or present to a high level. The protective indicators are scored in reverse, which is to say that lower scores indicate a higher level of risk (Pressman et al., 2018). A protective indicator is rated as 'low' if no risk-mitigating indicator characteristics are present, as 'moderate' if some risk-mitigating indicator characteristics are present, and as 'high' if clear risk-mitigating indicator characteristics are present (Pressman et al., 2018). It is important to stress here that the VERA-2R does not provide a numerical score for the ratings

(Pressman et al., 2018). However, for the purposes of this study we assigned the numerical scores '0', '1' and '2' to the ratings 'low', 'moderate' and 'high', respectively. Based on the assumption that the indicator characteristics will be cited in the judicial file if they are present, we decided to assign the numerical score '0' if the judicial file did not contain information about an indicator.

The VERA-2R also includes 11 additional indicators, which may contribute to a person's vulnerability to engage in future acts of violent extremism and terrorism, when combined with the presence of ideological, contextual, and motivational indicators identified in the VERA-2R (Pressman et al., 2018). These additional indicators are divided into three domains: Criminal History (CH), Personal History (PH) and Mental Disorder (MD). The scientific basis for each indicator is explained, along with the criteria for the two rating levels: not present or present. The rating 'not present' (0) corresponds to the absence of the additional indicator characteristics, while the rating 'present' (1) corresponds to the presence of the additional indicator characteristics (Pressman et al., 2018). If a judicial file did not contain information about an indicator, then the numerical score '0' was assigned.

After carefully considering the indicators, the assessor then assigns structured professional risk judgements to the VERA-2R domains (Pressman et al., 2018). Subsequently, a final structured professional risk judgement is made in terms of the likelihood of an individual engaging in ideologically motivated violence. The structured professional risk judgements are formulated in a risk narrative, as well as rated on a scale of low (0), moderate (1), and high (2) (Pressman et al., 2018). Furthermore, different risk scenarios are identified with a risk management strategy for each of these scenarios (Douglas et al., 2014; Logan, 2017).

Research Design

Our research design includes trained researchers and extensive judicial files, and therefore closely resembles VERA-2R assessments in practice.

While determining our research design, we took into account the use of comprehensive judicial files, which include information from a range of different sources, such as the police, the Public Prosecution Service, and forensic psychiatrists and psychologists. In order to ensure that the assessors did not assign different ratings to the indicators, as a result of relying on different sources that may have differing opinions about whether (and to what extent) the indicator characteristics are present, we decided to inform the second assessor which source the first assessor had used.

Security and Privacy

To ensure that there were no risks to the privacy of the subjects, we anonymized the data. Moreover, with regard to data protection, we stored the anonymized dataset in a secure digital environment, in order to protect the information against misuse, unauthorized access, disclosure and theft.

Statistical Analysis

The statistical analyses were carried out using IBM SPSS Statistics for Mac Version 25.0. The interrater reliability of the VERA-2R was examined by means of the intraclass correlation coefficient (ICC), using the two-way random effects model and absolute agreement type (Hallgren, 2012). The intraclass correlation coefficient (ICC) was chosen as the reliability index, because extant literature has showed that ICC is one of the most commonly-used statistics for interval variables (Hallgren, 2012). The ICC values were established for both the VERA-2R indicators and structured professional risk judgements. Furthermore, mean ICC values were determined by calculating the average amount of agreement over the VERA-2R indicators. Interpretation of ICCs were based on the critical values for single measures provided by Fleiss (1986): ICC < .40 = poor, .40 ≤ ICC < .60 = fair, .60 ≤ ICC < .75 = good and ICC ≥ .75 = excellent.

The interrater reliability of the additional indicators was examined by means of Cohen's kappa (κ). Furthermore, mean kappa values were determined by calculating the average amount of agreement over the additional indicators. Cohen's kappa (κ) was chosen as the reliability index for the additional indicators, because extant literature has showed that Cohen's kappa is one of the most commonly-used statistics for nominal variables (Hallgren, 2012). The kappa values were interpreted in accordance with the guidelines outlined by Landis and Koch (1977): κ ≤ .20 = slight, .20 < κ ≤ .40 = fair, .40 < κ ≤ .60 = moderate, .60 < κ ≤ .80 = good and .80 < κ ≤ .1.00 = excellent.

4. Results Interrater Reliability

Interrater Reliability of the VERA-2R Indicators

Table 1 shows the ICCs for the VERA-2R indicators. The indicators within the 'Beliefs, Attitudes and Ideology' domain (BA) all have good to excellent interrater reliability. The mean ICC value of the indicators is .79, which indicates excellent interrater reliability.

We also found excellent interrater reliability for most of the indicators (5 of 7) within the 'Social Context and Intention' domain (SCI). Furthermore, good interrater reliability was demonstrated for indicator 'SCI.3' (ICC = .72), while fair interrater reliability was demonstrated for indicator 'SCI.6' (ICC = .53). The average amount of agreement over the indicators can be classified as excellent (ICC = .82).

The indicators representing the 'History, Action and Capacity' domain (HAC) all have excellent interrater reliability, with the exception of indicator 'HAC.6' which was found to have fair interrater reliability (ICC = .51). The mean ICC value of the indicators is .85, which indicates excellent interrater reliability.

All the indicators within the 'Commitment and Motivation' domain (CM) have good to excellent interrater reliability, with the exception of indicator 'CM.5' which was found to have fair interrater reliability (ICC = .59). The average amount of agreement over the indicators can be classified as excellent (ICC = .78).

We also found excellent interrater reliability for the majority of the indicators (4 of 6) within the 'Protective and risk-mitigating indicators' domain (P). Furthermore, poor interrater reliability was demonstrated for indicator 'P.3' (ICC = .31), while fair interrater reliability was demonstrated for indicator 'P.6' (ICC = .53). The mean ICC value of the indicators is .73, which indicates good interrater reliability.

Overall, the average amount of agreement over the VERA-2R indicators can be classified as excellent (ICC = .79).

Table 1. *Interrater Reliability VERA-2R indicators*

Domain and indicator	N	ICC	95% CI
BA. Beliefs, Attitudes and Ideology			
BA.1 Commitment to ideology that justifies violence	30	.72***	.49 - .85
BA.2 Perceived grievances and/or perceived injustice	30	.81***	.64 - .91
BA.3 Dehumanization of designated targets associated with injustice	30	.72***	.49 - .86
BA.4 Rejection of democratic society and values	30	.64***	.37 - .81
BA.5 Expressed emotions in response to perceived injustice	30	.89***	.79 - .95
BA.6 Hostility to national identity	30	.90***	.80 - .95
BA.7 Lack of empathy and understanding for those outside one's own group	30	.85***	.71 - .93
Mean domain BA		.79	
SCI. Social Context and Intention			
SCI.1 Seeker, user or developer of violent extremist materials	30	.97***	.94 - .99
SCI.2 Target for attack identified (person, group, location)	30	.87***	.74 - .93
SCI.3 Personal contact with violent extremists (informal or social context)	30	.72***	.48 - .86
SCI.4 Expressed intention to commit acts of violent extremism	30	.83***	.66 - .92
SCI.5 Expressed willingness and/or preparation to die for a cause or belief	30	.83***	.67 - .91
SCI.6 Planning, preparation of acts of violent extremism	30	.53**	.22 - .74
SCI.7 Susceptibility to influence, control or indoctrination	30	.98***	.95 - .99
Mean domain SCI		.82	
HAC. History, Action and Capacity			
HAC.1 Early exposure to violence-promoting, militant ideology	30	.96***	.92 - .98
HAC.2 Network of family and friends involved in violent extremism	30	.78***	.59 - .89
HAC.3 Violent criminal history	30	.95***	.89 - .97
HAC.4 Strategic, paramilitary and/or explosives training	30	.94***	.88 - .97
HAC.5 Training in extremist ideology in own country or abroad	30	.95***	.90 - .98
HAC.6 Organizational skills and access to funding and sources of help	30	.51**	.19 - .74
Mean domain HAC		.85	

Domain and indicator	N	ICC	95% CI
CM. Commitment and Motivation			
CM.1 Motivated by perceived religious obligation and/or glorification	30	.74***	.52 - .87
CM.2 Motivated by criminal opportunism	30	.82***	.66 - .91
CM.3 Motivated by camaraderie, group belonging	30	.93***	.86 - .97
CM.4 Motivated by moral obligation, moral superiority	30	.82***	.65 - .91
CM.5 Motivated by excitement and adventure	30	.59***	.30 - .78
CM.6 Forced participation in violent extremism	30	.64***	.38 - .81
CM.7 Motivated by acquisition of status	30	.91***	.82 - .96
CM.8 Motivated by a search for meaning and significance in life	30	.81***	.64 - .91
Mean domain CM		.78	
P. Protective and risk-mitigating indicators			
P.1 Reinterpretation of the ideology	30	.94***	.88 - .97
P.2 Rejection of violence as a means to achieve goals	30	.76***	.55 - .88
P.3 Change in concept of the enemy	30	.31	.00 - .60
P.4 Participant in programmes against violent extremism	30	.94***	.87 - .97
P.5 Support from the community for non-violence	26	.89***	.78 - .94
P.6 Support from family members, other important persons for non-violence	30	.53**	.22 - .74
Mean domain P		.73	
Mean VERA-2R indicators		**.79**	

Note: ICC = Intraclass correlation coefficient and 95% CI = 95% confidence interval. The N of item 'P4' is 26, due to the fact that for some of the cases a suitable rating could not be assigned to the item.
*$p < .05$, **$p < .01$, ***$p < .001$*

Interrater Reliability of the Structured Professional Risk Judgements

Table 2 presents the ICCs for the structured professional risk judgements. We found good interrater reliability for the structured professional risk judgements across all the domains, with the exception of the domain 'Beliefs, Attitudes and Ideology' (BA), which was found to have excellent interrater reliability (ICC = .85). With respect to the final structured professional risk judgement, the results reveal an excellent level of agreement between the assessors (ICC = .81).

Table 2. *Interrater Reliability Structured Professional Risk Judgements*

Structured risk judgement	N	ICC	95% CI
Structured professional risk judgement domain 'Beliefs, Attitudes and Ideology'	30	.85***	.70 - .92
Structured professional risk judgement domain 'Social Context and Intention'	30	.74***	.52 - .87
Structured professional risk judgement domain 'History, Action and Capacity'	30	.70***	.46 - .84
Structured professional risk judgement domain 'Commitment and Motivation'	30	.74***	.53 - .87
Structured professional risk judgement domain 'Protective and risk-mitigating indicators'	30	.74***	.53 - .87
Final structured professional risk judgement	30	.81***	.64 - .91

Note: ICC = Intraclass correlation coefficient and 95% CI = 95% confidence interval
*$p < .05$, **$p < .01$, ***$p < .001$*

Interrater Reliability of the Additional Indicators

Table 3 shows the kappa values for the additional indicators. The additional indicators all have good to excellent interrater reliability, with the exception of indicator 'PH.3', which was found to have moderate interrater reliability (κ = .51). Furthermore, four indicators revealed a kappa coefficient of 1, which implies perfect interrater reliability. Overall, the average amount of agreement over the additional indicators can be classified as excellent (κ = .85).

Table 3. *Interrater Reliability Additional Indicators*

Additional indicators	N	Kappa (κ)	95% CI
CH. Criminal History			
CH.1 Client of the juvenile justice system/convicted for non-violent offence(s)	30	1.00***	1.00 - 1.00
CH.2 Non-compliance with conditions or supervision	13	.63*	-0.02 - 1.00
Mean domain CH		.82	
PH. Personal History			
PH.1 Violence in family	30	.90***	.71 - 1.00
PH.2 Problematic upbringing and/or placed in juvenile care	30	.80***	.59 - 1.00
PH.3 Problems with school and work	30	.51**	.20 - .82
Mean domain PH		.74	
MD. Mental Disorder			
MD.1 Personality disorder	30	.84***	.63 - 1.00
MD.2 Depressive disorder and/or suicide attempts	30	1.00***	1.00 - 1.00
MD.3 Psychotic and schizophrenic disorder	30	1.00***	1.00 - 1.00
MD.4 Autism spectrum disorder	30	1.00***	1.00 - 1.00
MD.5 Post-traumatic stress disorder	30	.87***	.62 - 1.00
MD.6 Substance use disorder	30	.75***	.49 - 1.00
Mean domain MD		.91	
Mean additional indicators		**.85**	

Note: κ = kappa value and 95% CI = 95% confidence interval. The N of item 'CH2' is 13, due to the fact that for some of the cases a suitable rating could not be assigned to the item.
**p < .05, **p < .01, ***p < .001*

5. Methods Intrarater Reliability

Assessors and Cases

The assessor was a Dutch researcher (female) with a bachelor degree in psychology and a master degree in criminology. The assessor was employed by the Netherlands Institute of Forensic Psychiatry and Psychology (NIFP) of the Dutch Ministry of Justice and Security, and took part in a two-day training course to obtain an in-depth understanding of the instrument and to acquire experience in applying the VERA-2R indicators and forming structured professional risk judgements.

To assess the intrarater reliability of the VERA-2R, the assessor rated a sample of terrorist cases twice, with an interval minimum of 6 months. In order to establish the minimum sample size, a-priori power analyses were performed: a-priori power analysis for the intraclass correlation coefficient (ICC) estimated that 28 cases[1] were required and a-priori power analysis for Cohen's kappa estimated that 33 cases[2] were required. Therefore, we selected 33 cases of terrorist offenders and assessed them on the basis of extensive judicial files. These files were provided by the Dutch Public Prosecution Service, and included a FMH assessment, a probation report, a transcript of the verdict, a police report, a criminal record and/or information from intelligence services.

The sample comprised 27 men and 6 women who were convicted of terrorist offences in the Netherlands between 2012 and 2019. The subjects' ages ranged from 15 - 59 at the time of their terrorist act (M_{age} = 26.12, *SD* = 9.74). 84.8% of the subjects were from a migrant background, 3.0% of the subjects had no migration background, and for 12.1% the background was unknown. Most of the subjects from a migrant background had parents who were born in Morocco or were born there themselves (46.4%), followed by persons with a Turkish background (17.9%). Consequently, the sample appears to be representative of the target population with respect to age, gender and migration background, since similar descriptive statistics were found in previous studies on Dutch jihadists (Bakker & De Bont, 2016; Weenink, 2019).

Materials

A detailed description of the VERA-2R is included in the material section of the interrater reliability study.

1 *We choose a minimum acceptable reliability of .40, since this indicates fair reliability. Furthermore, we choose an expected reliability of .75, since this corresponds to the interrater reliability we found for the continuous variables of the VERA-2R, and we expect that interrater reliability has a significant impact on intrarater reliability. Furthermore, we selected a power of .80, a significance level of .05, and two repetitions per subject.*

2 *We choose a minimum acceptable reliability of .40, since this indicates moderate reliability. Furthermore, we choose an expected reliability of .85, since this corresponds to the interrater reliability we found for the categorical variables of the VERA-2R, and we expect that interrater reliability has a significant impact on intrarater reliability. Furthermore, we selected a proportion of outcome of .50, a power of .80 and a significance level of .05.*

Research Design

As mentioned before, our research design includes a trained researcher and extensive judicial files, and therefore closely resembles VERA-2R assessments in practice.

In order to ensure that the assessor did not assign different ratings on T1 and T2 as a result of relying on different sources that may have differing opinions about whether (and to what extent) the indicator characteristics are present, we decided to inform the assessor at T2 which source she had used during T1.

With respect to the interval between T1 and T2, we took into account that a long interval increases the risk of changes in the observed individual, whereas a short interval increases the risk of recall bias. Since we evaluated the cases on the basis of static judicial files, we faced no risks of changes in the observed individual. Therefore, in order to be able to minimize the risk of recall bias, we chose a long interval of minimum 6 months.

Security and Privacy

To ensure that there were no risks to the privacy of the subjects, we anonymized the data. Moreover, with regard to data protection, we stored the anonymized dataset in a secure digital environment, in order to protect the information against misuse, unauthorized access, disclosure and theft.

Statistical Analysis

The statistical analyses were carried out using IBM SPSS Statistics for Mac Version 25.0. In order to establish the intrarater reliability of the VERA-2R, the intraclass correlation coefficient (ICC) was used (two-way mixed-effects model and absolute agreement type) (Hallgren, 2012). In line with the vision of Shrout and Fleiss (1979), the two-way mixed-effects model is chosen, as it is not reasonable to generalize the scores of one assessor to a larger population of assessors (Koo & Li, 2016). The ICC values were established for both the VERA-2R indicators and the structured professional risk judgements. Furthermore, mean ICC values were determined by calculating the average amount of agreement over the VERA-2R indicators. Interpretation of ICCs were based on the critical values for single measures provided by Fleiss (1986): ICC < .40 = poor, .40 ≤ ICC < .60 = fair, .60 ≤ ICC < .75 = good and ICC ≥ .75 = excellent.

The intrarater reliability of the additional indicators was examined by means of Cohen's kappa (κ). Furthermore, mean kappa values were determined by calculating the average amount of agreement over the additional indicators. The kappa values were interpreted in accordance with the guidelines outlined by Landis and Koch (1977): κ ≤ .20 = slight, .20 < κ ≤ .40 = fair, .40 < κ ≤ .60 = moderate, .60 < κ ≤ .80 = good and .80 < κ ≤ .1.00 = excellent.

6. Results Intrarater Reliability

Intrarater Reliability of the VERA-2R Indicators

Table 4 shows the results for the VERA-2R indicators. We found excellent intrarater reliability for all of the indicators within the 'Beliefs, Attitudes and Ideology' domain (BA), with 2 indicators revealing an ICC value of 1 (BA.3 and BA.5). The mean ICC value of the indicators is .96, which indicates excellent intrarater reliability.

The indicators representing the 'Social Context and Intention' domain (SCI) all have excellent intrarater reliability, with ICC values ranging from .81 (indicator SCI.5) to .94 (indicator SCI.6). The average amount of agreement over the indicators can be classified as excellent (ICC = .88).

All the indicators within the 'History, Action and Capacity' domain (HAC) have excellent intrarater reliability, with the exception of indicator 'HAC.4' which was found to have good intrarater reliability (ICC = .74). Furthermore, indicator 'HAC.1' revealed an ICC value of 1. The mean ICC value of the indicators is .91, which indicates excellent intrarater reliability.

Most of the indicators (6 of 8) within the 'Commitment and Motivation' domain (CM) have excellent intrarater reliability. Furthermore, fair intrarater reliability was demonstrated for indicator 'CM.5' (ICC = .50), while good intrarater reliability was demonstrated for indicator 'CM.8' (ICC = .60). Indicator 'CM.6' moreover revealed an ICC value of 1. The average amount of agreement over the indicators can be classified as excellent (ICC = .82).

We also found excellent intrarater reliability for all of the indicators within the 'Protective and risk-mitigating indicators' domain (P), with the exception of indicator 'P.6' which was found to have fair intrarater reliability (ICC = .57). The mean ICC value of the indicators is .80, which indicates excellent intrarater reliability.

Overall, the average amount of agreement over the VERA-2R indicators can be classified as excellent (ICC = .82).

Table 4. *Interrater Reliability VERA-2R Indicators*

Domain and indicator	N	ICC	95% CI
BA. Beliefs, Attitudes and Ideology			
BA.1 Commitment to ideology that justifies violence	30	.90***	.80 - .95
BA.2 Perceived grievances and/or perceived injustice	30	.98***	.96 - .99
BA.3 Dehumanization of designated targets associated with injustice	30	1.00	1.00 - 1.00
BA.4 Rejection of democratic society and values	30	.91***	.82 - .95
BA.5 Expressed emotions in response to perceived injustice	30	1.00	1.00 - 1.00
BA.6 Hostility to national identity	30	.94***	.89 - .97
BA.7 Lack of empathy and understanding for those outside one's own group	30	.96***	.92 - .98
Mean domain BA		.96	
SCI. Social Context and Intention			
SCI.1 Seeker, user or developer of violent extremist materials	30	.93***	.86 - .97
SCI.2 Target for attack identified (person, group, location)	30	.83***	.68 - .91
SCI.3 Personal contact with violent extremists (informal or social context)	30	.86***	.73 - .93
SCI.4 Expressed intention to commit acts of violent extremism	30	.91***	.83 - .96
SCI.5 Expressed willingness and/or preparation to die for a cause or belief	30	.81***	.65 - .90
SCI.6 Planning, preparation of acts of violent extremism	30	.94***	.89 - .97
SCI.7 Susceptibility to influence, control or indoctrination	30	.90***	.79 - .95
Mean domain SCI		.88	
HAC. History, Action and Capacity			
HAC.1 Early exposure to violence-promoting, militant ideology	30	1.00	1.00 - 1.00
HAC.2 Network of family and friends involved in violent extremism	30	.87***	.76 - .94
HAC.3 Violent criminal history	30	.98***	.95 - .99
HAC.4 Strategic, paramilitary and/or explosives training	30	.74***	.54 - .86
HAC.5 Training in extremist ideology in own country or abroad	30	.95***	.91 - .98
HAC.6 Organizational skills and access to funding and sources of help	30	.89***	.79 - .94
Mean domain HAC		.91	

Domain and indicator	N	ICC	95% CI
CM. Commitment and Motivation			
CM.1 Motivated by perceived religious obligation and/or glorification	30	.81***	.62 - .91
CM.2 Motivated by criminal opportunism	30	.95***	.90 - .97
CM.3 Motivated by camaraderie, group belonging	30	.92***	.84 - .96
CM.4 Motivated by moral obligation, moral superiority	30	.90***	.80 - .95
CM.5 Motivated by excitement and adventure	30	.50**	.21 - .72
CM.6 Forced participation in violent extremism	30	1.00	1.00 - 1.00
CM.7 Motivated by acquisition of status	30	.87***	.74 - .93
CM.8 Motivated by a search for meaning and significance in life	30	.60***	.33 - .79
Mean domain CM		.82	
P. Protective and risk-mitigating indicators			
P.1 Reinterpretation of the ideology	30	.80***	.64 - .90
P.2 Rejection of violence as a means to achieve goals	30	.82***	.67 - .91
P.3 Change in concept of the enemy	30	.87***	.75 - .93
P.4 Participant in programmes against violent extremism	19	.78***	.51 - .91
P.5 Support from the community for non-violence	30	.96***	.93 - .98
P.6 Support from family members, other important persons for non-violence	30	.57***	.28 - .76
Mean domain P		.80	
Mean VERA-2R indicators		**.82**	

Note: ICC = Intraclass correlation coefficient and 95% CI = 95% confidence interval. The N of item 'P4' is 19, due to the fact that for some of the cases a suitable rating could not be assigned to the item.
*$p < .05$, **$p < .01$, ***$p < .001$*

Intrarater Reliability of the Structured Professional Risk Judgements

Table 5 presents the ICCs for the structured professional risk judgements. We found good to excellent intrarater reliability for the structured professional risk judgements across all the domains, with ICC values ranging from .64 (Domain SCI) to .88 (Domain HAC). With respect to the final structured professional risk judgement, the results reveal an excellent level of agreement (ICC = .83).

Table 5. *Interrater Reliability Structured Professional Risk Judgements*

Structured risk judgement	N	ICC	95% CI
Structured professional risk judgement domain 'Beliefs, Attitudes and Ideology'	30	.86***	.74 - .93
Structured professional risk judgement domain 'Social Context and Intention'	30	.64***	.38 - .80
Structured professional risk judgement domain 'History, Action and Capacity'	30	.88***	.77 - .94
Structured professional risk judgement domain 'Commitment and Motivation'	30	.72***	.51 - .85
Structured professional risk judgement domain 'Protective and risk-mitigating indicators'	30	.86***	.74 - .93
Final structured professional risk judgement	30	.83***	.68 - .91

Note: ICC = Intraclass correlation coefficient and 95% CI = 95% confidence interval
*$p < .05$, **$p < .01$, ***$p < .001$*

Intrarater Reliability of the Additional Indicators

The results for the additional indicators are demonstrated in Table 6. The additional indicators all have good to excellent intrarater reliability, with the exception of indicator 'PH.3', which was found to have moderate intrarater reliability (κ = .59). Furthermore, four indicators revealed a kappa coefficient of 1, which implies perfect intrarater reliability. Overall, the average amount of agreement over the additional indicators can be classified as excellent (κ = .86).

Table 6. *Interrater Reliability Additional Indicators*

Additional indicators	N	Kappa (κ)	95% CI
CH. Criminal History			
CH.1 Client of the juvenile justice system/convicted for non-violent offence(s)	30	.93***	.80 – 1.00
CH.2 Non-compliance with conditions or supervision	11	.62***	-.04 – 1.00
Mean domain CH		.78	
PH. Personal History			
PH.1 Violence in family	30	.74***	.47 – 1.00
PH.2 Problematic upbringing and/or placed in juvenile care	30	.94***	.75 – 1.00
PH.3 Problems with school and work	30	.59**	.30 – .88
Mean domain PH		.76	
MD. Mental Disorder			
MD.1 Personality disorder	30	1.00***	1.00 - 1.00
MD.2 Depressive disorder and/or suicide attempts	30	1.00***	1.00 - 1.00
MD.3 Psychotic and schizophrenic disorder	30	1.00***	1.00 - 1.00
MD.4 Autism spectrum disorder	30	1.00***	1.00 - 1.00
MD.5 Post-traumatic stress disorder	30	.84***	.53 - 1.00
MD.6 Substance use disorder	30	.85***	.65 - 1.00
Mean domain MD		.95	
Mean additional indicators		**.86**	

Note: κ = kappa value and 95% CI = 95% confidence interval. The N of item 'CH2' is 13, due to the fact that for some of the cases a suitable rating could not be assigned to the item.
**p < .05, **p < .01, ***p < .001*

7. Discussion

Given that risk assessments have a significant role to play in the fight against violent extremism and terrorism (European Commission, 2020; United Nations, 2018), it is of importance that risk assessment tools for ideologically motivated violence provide reliable and valid risk assessments. The present study investigated the reliability of the VERA-2R, focusing on the interrater- and intrarater reliability. In accordance with our hypotheses, the results show that the reliability of the VERA-2R is good to excellent. This conclusion is first of all supported by the level of agreement on the indicators, which can be classified as excellent for both the interrater- and intrarater reliability. These results indicate that the indicators included in the risk assessment and their encoding rules are clearly expressed (Multon & Coleman, 2018) and can therefore be assessed in the same way by different assessors or during repeated risk assessments over time. In addition to the promising results regarding the indicators, we also found good to excellent interrater- and intrarater reliability for the structured professional risk judgements, with slightly higher results for the intrarater reliability in comparison to the interrater reliability. Although structured professional risk judgements have been shown to be vulnerable to subjective biases (Shepherd & Sullivan, 2017), these results nevertheless indicate that the way that assessors exercise their professional judgment to integrate, combine and weigh all the relevant information and data related to the indicators is stable across different assessors and during repeated risk assessments over time. This is a significant finding, first of all, because structured professional risk judgements serve as the basis for important decisions in the criminal justice process (Van der Heide et al., 2019), and, as such, should be independent of the observers or professional assessors (Jonsson & Svingby, 2007). Secondly, it provides certainty that the VERA-2R is able to assess the risk status of an individual in the same way over time, and therefore can be used to identify changes in risk. This is important, since the dynamic nature of the process of radicalization to ideologically motivated violence and vice versa requires repeated VERA-2R risk assessments (Pressman et al., 2018).

Our findings are in line with a previous reliability study of the VERA (Beardsley & Beech, 2013). However, as aforementioned, Beardsley and Beech (2013) used case vignettes, had a small sample size, and did not include a stringent training program. Although these limitations can impact research findings, they are nevertheless regularly reported within the field of reliability studies of structured professional risk assessment tools (e.g., Sutherland et al., 2012; Svalin, Mellgren, Torstensson Levander, & Levander, 2017; Vial, Assink, Stams, & Van der Put, 2019). Given that our research design included extensive judicial files, samples of at least 30 cases, and assessors who were trained in the use of the VERA-2R and had a bachelor and master degree in psychology and/or criminology, we were able to overcome these limitations and, in turn, reflect the practice of structured professional risk assessment as closely as possible. Therefore, our research findings provide an initial indication that the VERA-2R can produce high interrater- and intrarater reliability in practice. Ideally, we would like to verify this further by establishing whether the same level of consistency would be found if the VERA-2R assessment was carried out by trained professionals with experience in carrying out individual risk assessments and also included information obtained from a direct interview with the person concerned. Unfortunately, since it is virtually impossible to have experts assess the

same suspect of ideologically motivated violence twice, or by two different experts or by one single expert but repeated over time with static information, this is not feasible for both types of reliability.

While the reliability of the VERA-2R can be classified as good to excellent, the instrument contains some indicators that appear to be more difficult to assess in the same way by different assessors and/or during repeated risk assessments over time. First of all, for 3 indicators low levels of both interrater- and intrarater reliability were found. The interrater and intrarater reliability of the indicator 'Motivated by excitement and adventure' (CM.5) may be low due to the fact that the concepts of 'excitement' and 'adventure' are not clearly defined. As a result, it may be difficult to achieve high levels of agreement, both across different assessors and during repeated risk assessments over time. However, in order to be able to provide a clearer explanation and a well-founded recommendation for the improvement of the reliability values, it is necessary to interview professionals with extensive experience in conducting VERA-2R risk assessments. With regard to the indicator 'Support from family members (to relinquish the use of violence)' (P.6) one could argue that we found low interrater- and intrarater reliability due to the fact that it is difficult to determine whether the person concerned was favorably influenced by the support. Higher levels of reliability could be achieved if the indicator focused on whether the person concerned receives support instead of whether the person concerned is positively affected by this support. The interrater-and intrarater reliability of the indicator 'Problems with school and work' (PH.3) may be low, because the word 'problems' leaves room for subjective interpretation. As a result, this interpretation may differ from researcher to researcher, or from time to time. In order to increase the reliability values, one could seek to objectify the content of the indicator by replacing 'Problems with school and work' with 'School dropout and work-related dismissal'.

In addition to this, 3 indicators were found to have high levels of intrarater reliability, but low levels of interrater reliability. The interrater reliability of the indicator 'Planning or preparation of acts of violent extremism' (SCI.6), may be low due to the lack of clarity over how the indicator should be assessed if the person concerned is suspected or convicted of a crime that sought to prepare or facilitate a violent extremist or terrorist crime. This concerns specific types of crime, such as financing and incitement, which are not preceded by clear preparatory acts. In order to increase the interrater reliability of this indicator, clarification is required over how the indicator should be assessed with respect to these types of crime. With regard to the indicator 'Organizational skills and access to funding and sources of help' (HAC.6), one could argue that the indicator encompasses too many different concepts related to the ability to plan and execute violent extremist or terrorist acts. Higher interrater reliability may thus be achieved if the indicator were to focus on the organizational skills of the person concerned, with access to funding and sources of help as constituting examples from which this could be derived. The interrater reliability of the indicator 'Change in concept of the enemy' (P.3) may be low due to a lack of clarity over how the indicator should be assessed if the person concerned has no enemy. Given that enemy images are closely linked to grievances about perceived injustices, higher interrater reliability could be achieved if the indicator also focused on changes in grievances. Hereby it is good to mention that the recommendations regarding to the last two indicators have already been implemented in practice by incorporating it in the VERA-2R training.

Limitations and Recommendations

Although our findings undoubtedly provide valuable insights into the reliability of the VERA-2R, there are some limitations. The first limitation is that, although the research design simulates the practice of structured professional risk assessment as closely as possible, it needs to be defined as a research setting. In this study, VERA-2R risk assessments were carried out by researchers who were trained in the use of the VERA-2R on the basis of judicial files, while, in practice, VERA-2R risk assessments are carried out by professionals who have experience in undertaking individual risk assessments, preferably on the basis of judicial files and direct interviews with the person concerned. Since assessors must use their professional judgement to arrive at structured professional risk judgements (Pressman et al., 2018), the use of research assessors can be criticized on the grounds of their ability to form adequate structured professional risk judgements. However, we (partially) overcame this limitation by providing a two-day training course for the researchers, in which they acquired experience in forming structured professional risk judgements. With regard to the exclusion of interviews, it's important to state that the inclusion of personal interviews is not a requirement for the use of the VERA-2R. If the person concerned is absent or refuses to co-operate, VERA-2R risk assessments can and will be carried out without the information that would be obtained from a direct interview (Pressman et al., 2018).

A second limitation pertains to the fact that the results are based on a small sample size. Given that a larger sample size produces more reliable results with greater precision and power (Pallant, 2016), follow-up research should be carried out to determine if the results can be replicated in a larger sample.

Taking all this into consideration, we can conclude that the present study did not only clarify the reliability of the VERA-2R, it also showed how the reliability of a structured professional risk assessment tool can be investigated with trained assessors based on extensive judicial files. As with most research, the obtained knowledge can be deepened and strengthened by carrying out further research. In addition to this, it is also necessary to strengthen the empirical foundations of the VERA-2R. Due to both the limited access to (primary) data (Sageman, 2014) and the ethical barriers in conducting research on sensitive topics (Horgan, 2012), the evidence-base underpinning the risk-promoting and risk-mitigating indicators for ideologically motivated violence is scant at best (Sarma, 2017). In order to obtain a more evidence-based professional approach to conducting violent extremism and terrorism risk assessments, the European Database of Convicted Terrorists (EDT) was developed (Alberda et al., 2021). The EDT is based on judicial documents and contains personal and contextual information about convicted (and deceased) terrorists and violent extremists. By analyzing this data, reliable insights could be obtained into the underlying indicators that drive individuals' engagement, continuation or disengagement in violent extremism and terrorism. Subsequently, this would enable the validation of the VERA-2R indicators, as well as the identification of other relevant indicators vis-à-vis the risk of ideologically motivated violence. Furthermore, it is important to establish other psychometric properties of the VERA-2R, such as the discriminative validity, divergent validity and predictive validity, in order to be able to develop a more reliable and valid VERA tool. These aforementioned aspects will be investigated in follow-up research. Upon completion of this further research, we will be able to make well-founded recommendations on how to improve the VERA-2R, which, in turn, will lead to more accurate violent extremism and terrorism risk assessments and risk management strategies, and, most importantly, a safer society.

8. References

Alberda, D., Duits, N., Van den Bos, K., Ayanian, A. H., Zick, A., & Kempes, M. (2021). The European Database of Terrorist Offenders (EDT). *Perspectives on Terrorism, 15(2),* 77 - 99.

Bakker, E., & De Bont, R. (2016). Belgian and Dutch jihadist foreign fighters (2012-2015): Characteristics, motivations, and roles in the war in Syria and Iraq. *Small Wars & Insurgencies, 27(5),* 837 – 857. https://doi.org/10.1080/09592318.2016.1209806

Beardsley, N. L., & Beech, A. R. (2013). Applying the violent extremist risk assessment (VERA) to a sample of terrorist case studies. Journal of Aggression, *Conflict and Peace Research, 5(1),* 4 - 15. https://doi.org/10.1108/17596591311290713

Belur, J., Tompson, L., Thornton, A., & Simon, M. (2021). Interrater reliability in systematic review methodology: exploring variation in coder decision-making. *Sociological methods & research, 50(2),* 837 - 865.

Douglas, K. S., Hart, S. D., Webster, C. D., Belfrage, H., Guy, L. S., & Wilson, C. M. (2014). Historical-clinical-risk management-20, version 3 (HCR-20V3): Development and overview. *International Journal of Forensic Mental Health, 13(2),* 93 - 108. https://doi.org/10.1080/14999013.2014.906519

Duits, N., Rinne, T., & Van Leyenhorst, M. (2017). De risicoanalyse van gewelddadig extremisme in het strafrecht. *Sancties 2017/4,* 215 - 225.

European Commission. (2020). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on the EU Security Union Strategy.* COM(2020) 605 final. Retrieved from https://eur-lex.europa.eu/legalcontent/EN/TXT/PDF/ ?uri=CELEX:52020DC0605&

Europol (2022). *European Union Terrorism Situation and Trend report (TE-SAT) 2022.* Retrieved from https://www.europol.europa.eu/cms/sites/default/files/documents/Tesat_Report_2022_ 0.pdf

Field, A. P. (2005). *Discovering statistics using SPSS* (2nd ed.). London, England: Sage

Fleiss, J. L. (1986). *The design and analysis of clinical experiments.* New York, NY: Wiley.

Groth-Marnat, G. (2009). *Handbook of psychological assessment (5th ed.).* Hoboken, NJ: John Wiley & Sons.

Guy, L. S., Packer, I. K., & Warnken, W. (2012). Assessing risk of violence using structured professional judgment guidelines. *Journal of Forensic Psychology Practice, 12(3),* 270 - 283. https://doi.org/10.1080/15228932.2012.674471

Gyldmark, M., & Morrison, G. C. (2001). Demand for health care in Denmark: Results of a national sample survey using contingent valuation. *Social Science & Medicine, 53(8),* 1023 -1036. https://doi.org/10.1016/S0277-9536(00)00398-1

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8(1),* 23 - 34. https://doi.org/10.20982/tqmp.08.1.p023

Hart, S. D. (2009). Evidence-based assessment of risk for sexual violence. *Chapman Journal of Criminal Justice, 1(1),* 143 - 165.

Hart, S. D., & Logan, C. (2011). Formulation of violence risk using evidence-based assessments: The structured professional judgment approach. In P. Sturmey & M. McMurran (Eds.), *Forensic case formulation* (pp. 81 - 106). Chichester, England: Wiley Blackwell.

Hartling, L., Hamm, M., Milne, A., Vandermeer, B., Santaguida, P. L., Ansari, M., Dryden, D. M. (2012). *Validity and inter-rater reliability testing of quality assessment instruments.* Rockville, MD: Agency for Healthcare Research and Quality.

Helmus, L. M., & Thornton, D. (2015). Stability and predictive and incremental accuracy of the individual items of Static-99R and Static-2002R in predicting sexual recidivism: A meta-analysis. *Criminal Justice and Behavior, 42(9),* 917 - 937. https://doi.org/10.1177/0093854814568891

Holden, R. B. (2010). Face validity. In I. B. Weiner & W. E. Craighead (Eds.), *The corsini encyclopedia of psychology* (4th ed.) (pp. 637 - 638). Hoboken, NJ: Wiley.

Holmefur, M., Aarts, P., Hoare, B., & Krumlinde-Sundholm, L. (2009). Test-retest and alternate forms reliability of the assisting hand assessment. *Journal of rehabilitation medicine, 41(11),* 886 - 891.

Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports medicine, 30(1),* 1 - 15.

Horgan, J. (2012). Interviewing the terrorists: Reflections on fieldwork and implications for psychological research. *Behavioral Sciences of Terrorism and Political Aggression, 4(3),* 195 - 211. https://doi.org/10.1080/19434472.2011.594620

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2(2),* 130 - 144. https://doi.org/10.1016/j.edurev.2007.05.002

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine, 15(2),* 155 - 163.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33(2),* 363 - 374. https://doi.org/10.2307/2529786

Logan, C. (2017). Reporting structured professional judgement. In E. Bowen, S. Brown & D. Prescott (Eds.), *The forensic psychologist's report writing guide* (pp. 82 - 93). Chichester, UK: Wiley Blackwell.

Monahan, J. (2012). The individual risk assessment of terrorism. *Psychology, Public Policy, and Law, 18(2)*, 167 - 205. https://doi.org/10.1037/a0025792

Multon, K. D., & Coleman, J. S. M. (2018). Inter-rater reliability. In B. B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation* (pp. 863 - 865). Thousand Oaks, Canada: SAGE Publications.

Nationaal Coördinator Terrorismebestrijding en Veiligheid [National Coordinator for Security and Counterterrorism] (2016). Nationale contraterrorisme strategie 2016-2020 [National counterterrorism strategy 2016-2020]. Retrieved from https://www.nctv.nl/binaries/nctv/documenten/rapporten/2016/07/11/nationale-contraterroris mestrategie-2016-2020/CT-strategie+2016-2020.pdf

Pallant, J. (2016). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (6th ed.). London, England: McGraw-Hill Education.

Pressman, D. E. (2009). *Risk assessment decisions for violent political extremism.* Ottawa, Canada: Public Safety Canada.

Pressman, D. E., & Duits, N. (2019). De VERA à VERA-2R: Nouvelles avancées dans l'évaluation du risque d'extrémisme politique violent [From VERA to VERA-2R: New advances in assessing the risk of violent political extremism]. *Les Cahiers de la Sécurité et de la Justice, 46*, 57 - 71.

Pressman, D. E., Duits, N., Rinne, T., & Flockton, J. S. (2018). *VERA-2R Violent Extremism Risk Assessment - Version 2 Revised: A structured professional approach. Utrecht, The Netherlands: The Netherlands Institute of Forensic Psychiatry and Psychology (NIFP)*

Pressman, D. E., & Flockton, J. S. (2012). Calibrating risk for violent political extremists and terrorists: The VERA 2 structured assessment. *The British Journal of Forensic Practice, 14(4)*, 237 - 251. https://doi.org/10.1108/14636641211283057

Roychowdhury, A., & Adshead, G. (2014). Violence risk assessment as a medical intervention: Ethical tensions. *Psychiatric Bulletin, 38(2)*, 75 - 82. https://doi.org/10.1192/pb.bp.113.043315

Sageman, M. (2014). The stagnation in terrorism research. *Terrorism and Political Violence, 26(4)*, 565 - 580. https://doi.org/10.1080/09546553.2014.895649

Sarma, K. M. (2017). Risk assessment and the prevention of radicalization from nonviolence into terrorism. *American Psychologist, 72(3)*, 278 - 288. https://doi.org/10.1037/amp0000121

Schmid, A. (2013). Radicalisation, de-radicalisation, counter-radicalisation: A conceptual discussion and literature review. *ICCT Research Paper, 97(1)*, 1 - 22. https://doi: 10.19165/2013.1.02

Shepherd, S. M., & Sullivan, D. (2017). Covert and implicit influences on the interpretation of violence risk instruments. *Psychiatry, Psychology and Law, 24(2),* 292- 301. https://doi.org/10.1080/13218719.2016.1197817

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin, 86(2),* 420 - 428.

Striegher, J. L. (2015). Violent-extremism: An examination of a definitional dilemma. In *The Proceedings of [the] 8th Australian Security and Intelligence Conference* (pp. 75 - 86). Perth, Australia: ECU Security Research Institute, Edith Cowan University Joondalup Campus Western Australia. https://doi.org/10.4225/75/57a945ddd3352

Sutherland, A. A., Johnstone, L., Davidson, K. M., Hart, S. D., Cooke, D. J., Kropp, P. R., ... Stocks, R. (2012). Sexual violence risk assessment: An investigation of the interrater reliability of professional judgments made using the Risk for Sexual Violence Protocol. *International Journal of Forensic Mental Health, 11(2),* 119 - 133. https://doi.org/10.1080/14999013.2012.690020

Svalin, K., Mellgren, C., Torstensson Levander, M., & Levander, S. (2017). The inter-rater reliability of violence risk assessment tools used by police employees in Swedish police settings. *Nordisk Politiforskning 4(1),* 9 - 28. https://doi.org/10.18261/ISSN.1894-8693-2017-01-03

United Nations. (2018). *Draft resolution submitted by the President of the General Assembly - The United Nations Global Counter-Terrorism Strategy Review.* Seventy second session. A/72/L.62. Retrieved from https://undocs.org/pdf?symbol=en/A/RES/72/284

Van der Heide, L., Van der Zwan, M., & Van Leyenhorst, M. (2019). *The Practitioner's guide to the galaxy - A comparison of risk assessment tools for violent extremism.* Geraadpleegd van https://icct.nl/app/uploads/2019/09/29Aug19_Formatted_ThePractitionersGuideto theGalaxy-2.pdf

Vial, A., Assink, M., Stams, G. J. J., & Van der Put, C. (2019). Safety and risk assessment in child welfare: A reliability study using multiple measures. *Journal of Child and Family Studies, 28(12),* 3533 - 3544.

Weenink, A. W. (2019). Adversity, criminality, and mental health problems in Jihadis in Dutch police files. *Perspectives on Terrorism, 13(5),* 130 - 142.

Yang, M., Wong, S. C. P., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin, 136(5),* 740 - 767. https://doi.org/10.1037/a0020473